nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-022-01250-0

Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro

Bingjie Zhang[1,2,4], Avi Srivastava[1,2,4], Eleni Mimitou[3], Tim Stuart[1,2], Ivan Raimondi[3], Yuhan Hao[1,2], Peter Smibert[3] and Rahul Satija[1,2] ✉

Technologies that profile chromatin modifications at single-cell resolution offer enormous promise for functional genomic characterization, but the sparsity of the measurements and integrating multiple binding maps represent substantial challenges. Here we introduce single-cell (sc)CUT&Tag-pro, a multimodal assay for profiling protein–DNA interactions coupled with the abundance of surface proteins in single cells. In addition, we introduce single-cell ChromHMM, which integrates data from multiple experiments to infer and annotate chromatin states based on combinatorial histone modification patterns. We apply these tools to perform an integrated analysis across nine different molecular modalities in circulating human immune cells. We demonstrate how these two approaches can characterize dynamic changes in the function of individual genomic elements across both discrete cell states and continuous developmental trajectories, nominate associated motifs and regulators that establish chromatin states and identify extensive and cell-type-specific regulatory priming. Finally, we demonstrate how our integrated reference can serve as a scaffold to map and improve the interpretation of additional scCUT&Tag datasets.

Technologies that enable unsupervised transcriptomic profiling at single-cell resolution (i.e., single-cell RNA sequencing (scRNA-seq)) represent powerful tools not only for the discovery of cell types and states but also to reveal the structure and function of transcriptional regulatory networks[1–6]. In the same way, new methods for single-cell chromatin profiling, such as scATAC-seq, can lead to the identification and characterization of individual genomic regulatory regions and the exploration of their variation in a heterogeneous population[7,8]. Assays that profile chromatin accessibility exhibit a largely binarized phenotype, partitioning genomic regions into accessible or inaccessible elements in different tissue and cellular contexts. As a complement to accessibility measurements, genome-wide histone modification profiles offer an exciting opportunity to segment the genome into a more nuanced set of functional elements[9–12]. The 'histone code hypothesis' proposes that the combinatorial presence of multiple histone modifications marking a genomic region serves as a proxy for its regulatory function[13,14].

Techniques like chromatin immunoprecipitation sequencing (ChIP-seq) and cleavage under targets and tagmentation (CUT&Tag) have been widely applied to measure the binding profiles of multiple transcription factors and histone modifications[15,16]. In addition, tailored computational tools such as ChromHMM[17] and Segway[10] can integrate the combinatorial binding patterns defined by each mark and assign individual genomic elements to a set of learned functional 'states'. Recently, multiple studies have demonstrated the ability to measure CUT&Tag profiles in single cells[16,18–21]. This represents an opportunity to not only identify regions that exhibit cell-type-specific accessibility but also highlight elements whose acquisition of activating, repressive or heterochromatic signatures varies within a heterogeneous population.

These advances lay out an exciting challenge for single-cell genomics: can the binding profiles of multiple histone modifications be used to infer the function of any individual genomic element within a single cell? Although the availability of single-cell CUT&Tag (scCUT&Tag) takes an important step toward this vision, two key challenges remain. First, single-cell chromatin data remain extremely sparse, particularly due to the inherent challenges associated with limited DNA input molecules. Second, scCUT&Tag experiments typically measure only a single histone modification profile in a single cell. Although informative, only the integrated combination of multiple histone modifications can be used to annotate functional state, but no current technology can robustly and simultaneously profile many histone modifications in the same cell, particularly when these marks may exhibit overlapping binding patterns. Recent pioneering advances enable paired measurements of cellular transcriptomes and individual CUT&Tag profiles using custom combinatorial indexing-based workflows[22,23] and demonstrate how multimodal technologies can facilitate integrative analysis.

Here we present two techniques aimed to address these key challenges. We first introduce scCUT&Tag with cell surface proteins (scCUT&Tag-pro), a multimodal single-cell technology, that enables simultaneous profiling of an individual CUT&Tag profile with surface protein abundances and is compatible with the widely used 10x Genomics Chromium system. Second, we introduce a downstream analysis strategy for multimodal datasets: single-cell ChromHMM (scChromHMM). Our method first integrates data from multiple scCUT&Tag-pro experiments together into a common manifold, computationally generating coassay profiles for six histone modifications within individual cells. These measurements are used as input to a single-cell extension of the ChromHMM algorithm to return chromatin state annotations for each 200-bp genomic window at single-cell resolution.

We apply our tools and technologies to generate 64,876 scCUT&Tag-pro profiles (Supplementary Table 1) and combine this with existing datasets to create an integrated multimodal atlas of human peripheral blood mononuclear cells (PBMCs) that encompasses nine modalities spanning the central dogma. Together, these

[1]New York Genome Center, New York, NY, USA. [2]Center for Genomics and Systems Biology, New York University, New York, NY, USA. [3]Technology Innovation Lab, New York Genome Center, New York, NY, USA. [4]These authors contributed equally: Bingjie Zhang, Avi Srivastava.
✉e-mail: rsatija@nygenome.org

technologies and analytical tools allow us to explore heterogeneity in genome function at single-cell resolution, identify sequence elements whose presence accompanies changes in chromatin state and project new datasets onto our multimodal reference. We propose that these tools will help to facilitate the analysis and interpretation of chromatin state heterogeneity in single cells and lead to a better understanding of its role in establishing and regulating cellular identity.

Results

Simultaneous profiling of surface proteins and scCUT&Tag. The combinatorial pattern of post-translational modifications present on histone proteins correlates with the chromatin structure and regulatory potential of a genomic locus[13,14,24,25]. We reasoned that multimodal single-cell technologies could help address challenges in scCUT&Tag analysis by coupling robust single-cell measurements of one modality with more sparse measurements from another. For example, CITE-seq[26,27] and ASAP-seq[28,29] pair simultaneous measurements of highly expressed and well-characterized cell surface proteins with unsupervised but sparser transcriptomic measurements or chromatin accessibility profiles. We and others have recently demonstrated that the protein information in CITE-seq is highly informative in determining cellular state[30–32], whereas the paired RNA sequencing (RNA-seq) or assay for transposase-accessible chromatin sequencing (ATAC-seq) measurements enable the characterization of gene regulatory networks. ASAP-seq and integrated cellular indexing of chromatin landscape and epitopes sequencing are enabled by recent optimizations in cell fixation and permeabilization that enable ATAC-seq to be performed on whole cells instead of individual nuclei[33].

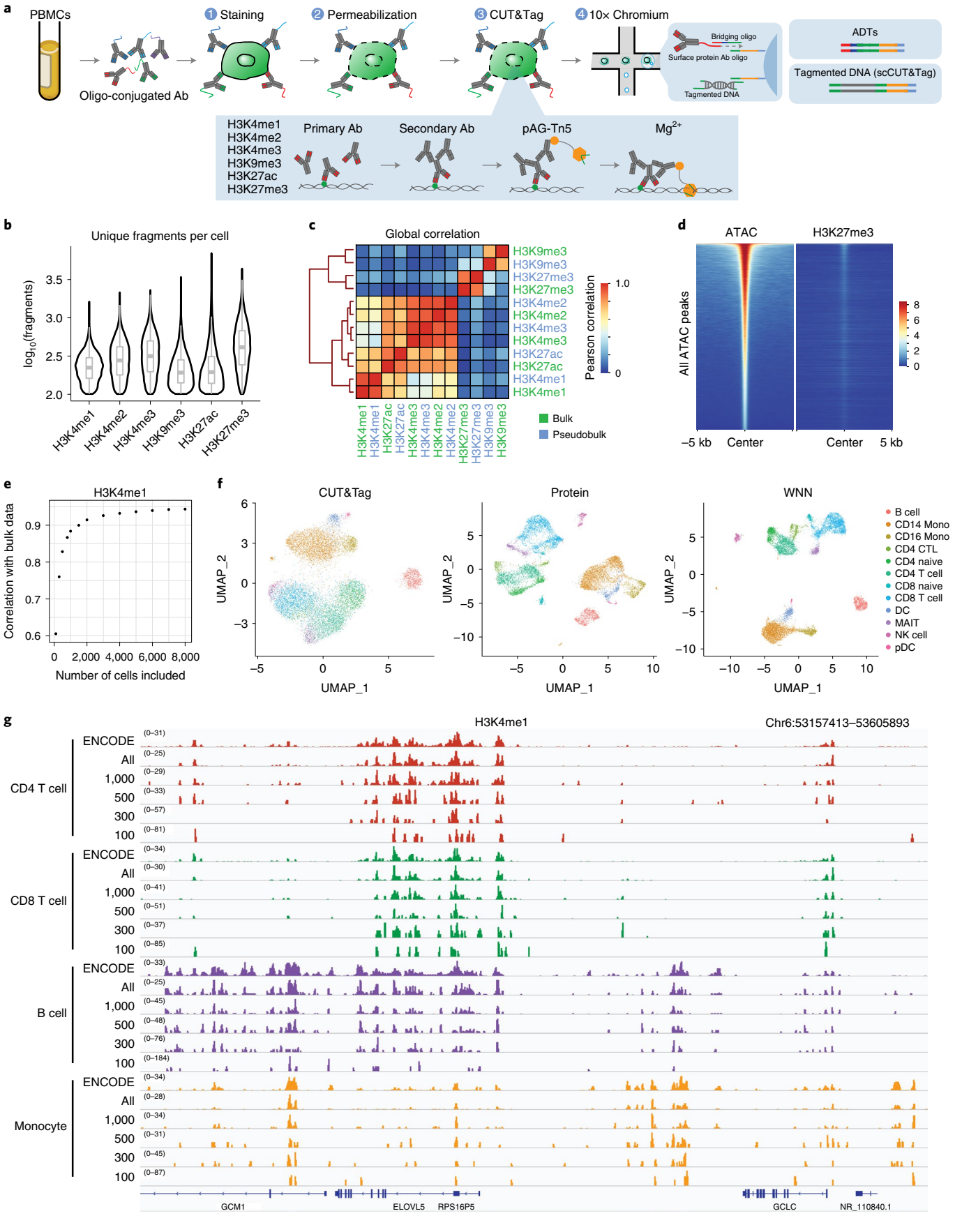
We propose that a similar multimodal strategy could be applied for genome-wide profiling of protein–DNA interactions as well. We developed scCUT&Tag-pro, an assay that performs CUT&Tag assays[16] on whole cells while simultaneously measuring surface protein levels (Fig. 1a and Methods). Inspired by ASAP-seq[28], our fixation and permeabilization conditions retain the cell membrane and associated proteins, enabling us to simultaneously measure cellular immunophenotypes (Fig. 1a). Briefly, cells are first stained with a panel of commercially available oligo-conjugated antibodies. Subsequently, monovalent Fab fragment is added to block the potential binding of proteinAG to these antibodies. Cells are then lightly fixed with 0.1% formaldehyde and permeabilized with an isotonic lysis buffer. We found that removing digitonin from all buffers substantially alleviated cell 'clumping' issues that have been previously described[18] without diminishing the efficiency of tagmentation (Supplementary Fig. 1a). We use an antibody against a histone modification of interest to direct the proteinAG-Tn5 fusion protein to marked genomic sites. Tagmented cells can then be used as input for library preparation using the 10x Genomics scATAC-seq kit, followed by sequencing of both CUT&Tag as well as antibody-derived tag (ADT) libraries. We note that scCUT&Tag-pro is also compatible with antibody-mediated cell hashing[34], allowing us to increase the throughput of each experiment.

In this study, we focused on acquiring scCUT&Tag-pro data from healthy human PBMCs and aimed to study the genome-wide localization patterns for six histone modifications (H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K27me3 and H3K9me3). In each experiment (Supplementary Table 1), we profiled a single histone modification, alongside an optimized panel of 173 antibodies from the BioLegend TotalSeq-A catalog (Supplementary Table 2). In total, we collected scCUT&Tag-pro from 64,876 individual cells (Supplementary Table 3).

We first evaluated the sensitivity and specificity of our CUT&Tag profiles. For example, in our H3K27me3 dataset, we observed, on average, 501 unique fragments per cell with an 84.1% uniquely mapping rate (Supplementary Fig. 1b), comparable to but slightly below a recent pioneering study[19] introducing scCUT&Tag on single nuclei from the same system (802 average unique fragments per cell; Fig. 1b and Supplementary Fig. 1c). For all marks, we observed that a pseudobulk profile generated from our single-cell data closely mirrored simultaneously generated bulk CUT&Tag profiles from PBMCs exhibited minimal overlap with accessible regions identified in a scATAC-seq dataset[28] (Fig. 1c,d). Through downsampling analysis, we found that the quantitative accuracy of each histone modification pseudobulk profile was dependent on the number of included single cells but began to saturate between 500 and 1,000 cells (Fig. 1e). Moreover, we found that unsupervised clustering of each scCUT&Tag dataset using a latent semantic indexing (LSI)-based dimensional reduction workflow[57] was capable of identifying major cell groups that define the human immune system (Fig. 1f) but did not recapitulate the high-resolution identification of cell states as observed in scRNA-seq or CITE-seq data[30].

Surface protein levels facilitate integration across experiments. We therefore examined the ADTs for 173 cell surface proteins, measured simultaneously in each experiment (Supplementary Fig. 1d). For example, in the H3K4me1 dataset, ADT expression patterns were fully concordant with cell types determined from chromatin-based clustering (i.e., uniform expression of CD3 in T cells, CD14 in monocytes and CD19 on B cells; Supplementary Fig. 1e). However, the protein modality revealed additional substructure, including mutual exclusivity of CD4 and CD8 in T cells, a division of monocytes into classical $CD14^+$ and nonclassical $CD16^+$ subgroups and canonical marker expression denoting mucosal-associated invariant T cells (Fig. 1f and Supplementary Fig. 1e). We used our recently introduced weighted nearest neighbor (WNN) analysis[30] (Methods) to simultaneously cluster scCUT&Tag-pro based on a weighted combination of modalities (Fig. 1f and Supplementary Fig. 2a,b). For each experiment, we found that the protein modality substantially enhanced our ability to separate distinct cell types, echoing our previous findings in CITE-seq and ASAP-seq analysis[26,28] (Supplementary Fig. 2c).

Fig. 1 | scCUT&Tag-pro enables simultaneous profiling of CUT&Tag and protein levels. a, Schematic of experimental workflow, which is compatible with the 10x Chromium system. **b**, Distribution of unique fragments obtained per cell for six histone modifications, profiled in separate PBMC scCUT&Tag-pro experiments (left to right: $n = 12{,}770$, 9,575, 10,386, 8,304, 15,609 and 8,232 cells). The center, bounds and whiskers of the boxplot show median, quartiles and data points that lie within 1.5× interquartile range of the lower and upper quartiles, respectively. Data beyond the end of the whiskers are plotted individually. **c**, Pseudobulk profiles of scCUT&Tag are well correlated with bulk CUT&Tag profiles of human PBMCs. **d**, Tornado plots of genomic regions ordered by chromatin accessibility. We observe no enrichment of H3K27me3 in accessible regions, indicating minimal open chromatin bias. **e**, Relationship between the number of cells included in a pseudobulk CUT&Tag profile of human PBMCs and the Pearson correlation with a bulk experiment. **f**, Uniform manifold approximation and projection (UMAP) visualizations of 12,770 single cells profiled with H3K4me1 scCUT&Tag-pro and clustered on the basis of CUT&Tag profiles, cell surface protein levels and WNN analysis, which combines both modalities. Cluster labels are derived from WNN analysis. **g**, Comparing pseudobulk CUT&Tag-pro profiles with ChIP-seq data from ENCODE. Including all cells assigned to each cell type results in pseudobulk tracks that closely mirror ENCODE profiles. However, even when downsampling to 300 cells per cluster, cell-type-specific patterns can still be observed. Ab, antibody; CTL, cytotoxic T lymphocyte; DC, dendritic cell; GCLC, glutamate-cysteine ligase catalytic subunit; MAIT, mucosal-associated invariant T cell; Mono, monocyte; NK, natural killer; pDC, plasmacytoid dendritic cell.


a
PBMCs
Oligo-conjugated Ab
1 Staining
2 Permeabilization
3 CUT&Tag
4 10× Chromium
Bridging oligo
Surface protein Ab oligo
Tagmented DNA
ADTs
Tagmented DNA (scCUT&Tag)
H3K4me1
H3K4me2
H3K4me3
H3K9me3
H3K27ac
H3K27me3
Primary Ab
Secondary Ab
pAG-Tn5
Mg2+
b
Unique fragments per cell
log10(fragments)
H3K4me1
H3K4me2
H3K4me3
H3K9me3
H3K27ac
H3K27me3
c
Global correlation
Pearson correlation
Bulk
Pseudobulk
d
ATAC
H3K27me3
All ATAC peaks
–5 kb
Center
5 kb
e
H3K4me1
Correlation with bulk data
Number of cells included
f
CUT&Tag
Protein
WNN
UMAP_1
UMAP_2
B cell
CD14 Mono
CD16 Mono
CD4 CTL
CD4 naive
CD4 T cell
CD8 naive
CD8 T cell
DC
MAIT
NK cell
pDC
g
H3K4me1
Chr6:53157413–53605893
CD4 T cell
CD8 T cell
B cell
Monocyte
ENCODE
All
GCM1
ELOVL5
RPS16P5
GCLC
NR_110840.1

We compared the results of our single-cell analysis with FACS-sorted H3K4me1 ChIP-seq profiles from ENCODE[35] (Methods). We observed that pseudobulk profiles generated from our WNN-derived scCUT&Tag-pro clusters clearly recapitulated the cell-type-specific ENCODE profiles (Fig. 1g). We also profiled biological replicates for one of the histone modifications (H3K4me1) and observed that these profiles were reproducible across replicates as well (Supplementary Fig. 2d). Moreover, downsampling our single-cell clusters to as low as 300 cells per profile still retained substantial cell type specificity in the ensuing pseudobulk tracks (Fig. 1g). Taken together, these analyses demonstrate that scCUT&Tag-pro datasets are comparable in sensitivity to existing technologies, exhibit low background, contain sufficient information to identify granular cell types and states and can be used to generate pseudobulk profiles that reflect high-quality bulk data when a sufficient number of cells are included.

As an alternative to fully unsupervised analysis, we found that single-cell protein measurements also facilitated high-resolution supervised annotation using reference datasets. For example, we recently introduced a carefully annotated reference dataset of 58 major and minor cell types and states in the circulating human immune system using a CITE-seq dataset of 161,764 cells and 228 surface proteins[30]. As our protein panel in this study largely overlaps with the reference dataset, we 'mapped' the cells from each scCUT&Tag-pro experiment onto our reference dataset (Methods) and repeated the procedure for our published ASAP-seq dataset[28] as well. Performing query-to-reference mapping using shared protein markers enables joint visualization of all datasets (Fig. 2a,b) and also transfers a consistent set of annotations to each cell at multiple levels of resolution. This enabled us to partition each dataset into broad (level 1), granular (level 2) and fine-grained (level 3) classifications (Supplementary Fig. 3a). We also used this approach to infer a unified pseudotime trajectory (Methods) that modeled CD8 T cell transitions from naive to effector states and confirmed the accuracy and resolution of our trajectory by comparing the developmental dynamics of key markers across experiments (Fig. 2c and Supplementary Fig. 3b).

Our reference-mapping workflow enabled us to explore the relationships between nine different modalities by harmonizing them into a common space and providing a consistent set of annotations (Fig. 2b). The modalities spanned the central dogma and ranged from measurements of chromatin accessibility, protein–DNA interactions (six targets), gene expression and protein abundances. Although the individual modalities were collected in separate experiments, each originated from a multimodal technology (ASAP-seq, scCUT&Tag-pro and CITE-seq) that was paired with a large and comprehensive panel of surface protein measurements to facilitate integration. For example, in Fig. 2d, we visualized data from each modality collected at the CD8A locus. As expected, we found robust gene and protein expression (as measured by CITE-seq), enriched chromatin accessibility (as measured by ASAP-seq) and the presence of activating histone modifications (as measured by scCUT&Tag-pro) in CD8 T cells. In myeloid cell types, the locus was characterized by inaccessible chromatin and the presence of repressive histone modifications, which resulted in undetectable expression of CD8 RNA and protein.

scChromHMM characterizes chromatin state at single-cell resolution. We next aimed to use our dataset to explore heterogeneity in chromatin state within the human immune system. We first considered an analysis strategy exclusively based on tools that have been previously developed for bulk chromatin analysis and provide 25 pseudobulk CUT&Tag tracks (based on level 2 reference-derived annotations) as input. For example, ChromHMM[12,17] trains a hidden Markov model (HMM) on the concatenated histone modification profiles for all cell types, and applies the Baum–Welch training algorithm[36] to identify a set of possible hidden 'chromatin states'. ChromHMM then applies the forward–backward algorithm[36] individually on each pseudobulk track to calculate the posterior probability that each 200-bp genomic window in each cell type is assigned to a particular chromatin state.

After running the Baum–Welch step (Methods), ChromHMM obtained 12 states (Fig. 3a), which could be broadly grouped into active promoter (enriched for H3K4me3/H3K4me2/H3K27ac), active enhancer (enriched for H3K4me1/H3K27ac), repressed (enriched for H3K27me3) and heterochromatic (enriched for H3K9me3) states. We also observed more subtle heterogeneity in chromatin state within these broad categories, likely corresponding to strong/weak states as has been previously described[17]. We identified a very similar set of chromatin states when running the Baum–Welch on pseudobulk tracks derived from level 3 annotations (Supplementary Fig. 4a,b). These results suggest that the Baum–Welch step of ChromHMM can be successfully run on pseudobulk tracks generated from scCUT&Tag-pro, returning a functionally interpretable set of chromatin states.

However, when attempting to associate these states with individual genomic elements in each cell type, we identified a limitation associated with the application of bulk methods to single-cell datasets. In particular, we found that in contrast to the initial Baum–Welch step for state identification, the final outputs of ChromHMM were highly dependent on the level of granularity used when generating pseudobulk profiles. For example, when we considered genomic windows assigned to enhancer states in $CD8^+$ T cells (level 1), 16.4% were not annotated as enhancers in any of the more granular T cell subsets when we repeated the analysis using a higher-resolution subset of cell labels (level 2). Especially when analyzing developmental tissues, or systems characterized by continuous sources of heterogeneity, it may be detrimental to condition all downstream analyses on a fixed set of discrete cell type labels[37].

We therefore devised an alternative approach, scChromHMM, where chromatin states are assigned at single-cell resolution. To do so, we first generated single-cell profiles with simultaneous measurements of six histone marks. Although it is not currently feasible to experimentally generate these data, we used our previously described anchoring workflow[38] as a computational alternative. We recently demonstrated the use of this workflow to 'transfer' modalities across experiments[38], for example, using a CITE-seq reference to accurately and robustly predict cell surface protein levels in an scRNA-seq dataset of human bone marrow. We applied a similar procedure in this study (Methods) to interpolate 20,000 single-cell profiles, each of which consisted of quantitative and genome-wide profiles for all six histone modifications and chromatin accessibility, alongside RNA and protein measurements. As in our previous studies, interpolated profiles represented a weighted average of 'anchor' cells, where the identification of anchors denoted a 'matching' of biological states across experiments based on shared protein expression patterns.

We assessed the accuracy of our interpolated profiles by comparing them to the original measurements that were obtained in separate experiments. For example, we computed the pseudobulk profile for H3K4me1 measurements based on our interpolated profiles, or the original measured values, and observed high quantitative concordance (Fig. 3b; $R = 0.95$). For each cell type, pseudobulk profiles of interpolated measurements clustered specifically with the pseudobulk profiles from the original measurements (Fig. 3c; additional histone marks are shown in Supplementary Fig. 5a, and cross-comparison across marks is show in Supplementary Fig. 5b), demonstrating that our procedure retains cell-type-specific variation. Finally, we repeated the interpolation procedure to generate an independent set of 20,000 profiles and observed high reproducibility ($R = 0.98$) between runs (Supplementary Fig. 5c). Although our procedure cannot capture stochastic fluctuations, it does represent

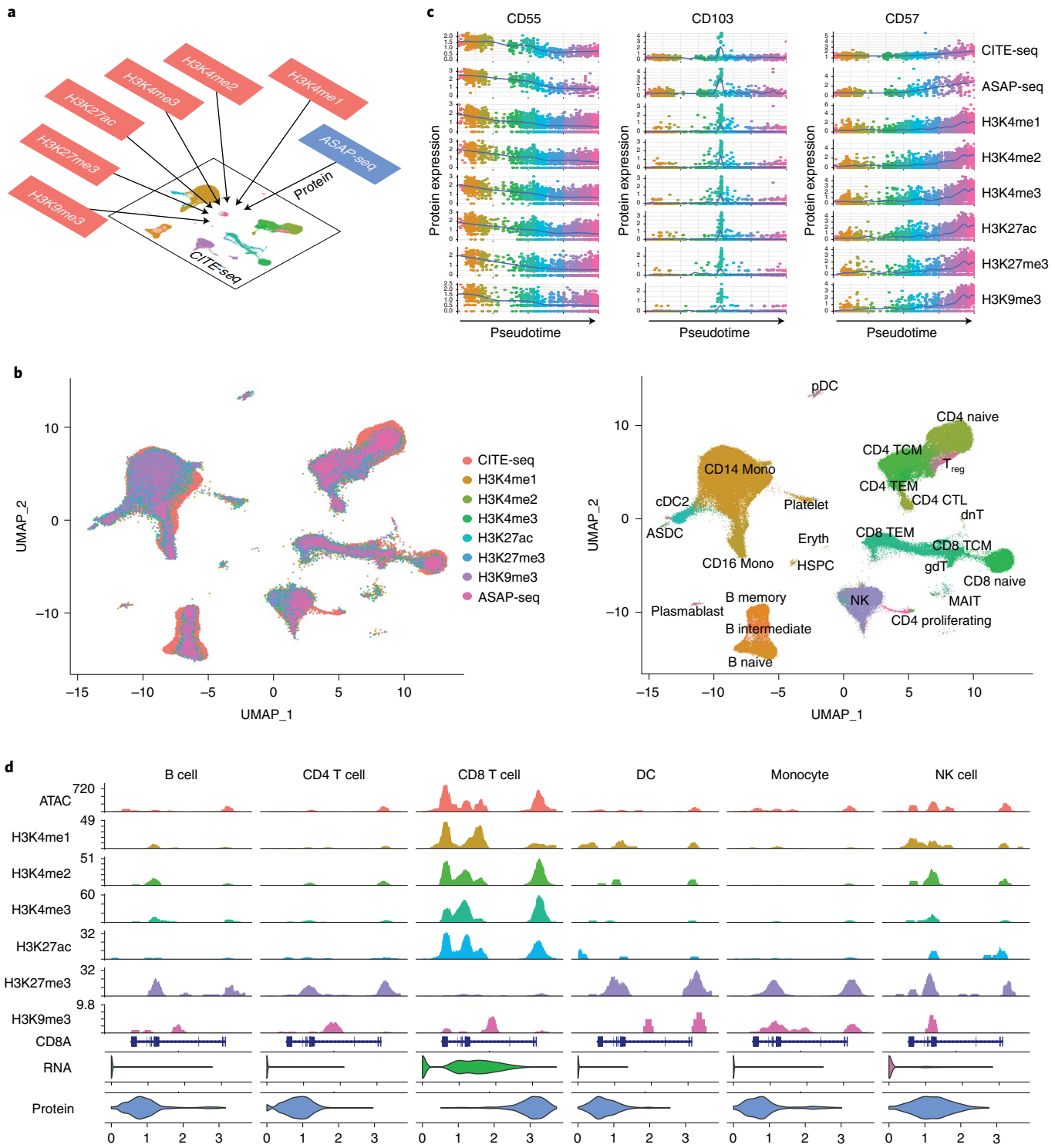


Fig. 2 | Protein measurements facilitate integrated analysis across modalities. a, Schematic workflow for integrated analysis. Datasets produced by scCUT&Tag-pro, ASAP-seq and CITE-seq are integrated together on the basis of a shared panel of cell surface protein measurements. **b**, Left: UMAP visualization of 230,597 total cells projected onto the reference dataset from Hao et al.[30]. Right: In addition to a harmonized visualization, cells from all experiments are annotated with a unified set of labels. **c**, We learned a unified pseudotime trajectory based on all experiments, representing the CD8 T cell transition from naive to effector states. We observe identical molecular dynamics for naive (CD55), memory (CD103) and effector (CD57) markers across all experiments, demonstrating that integrative analysis accurately identifies cells in matched biological states across experiments. **d**, Visualization of nine molecular modalities at the CD8A locus in B cell, CD4 T cell, CD8 T cell, dendritic cell, monocyte and NK cell groups. ASDC, AXL$^+$ dendridic cell; cDC2, conventional dendridic cell 2; dnT, double-negative T cell; Eryth, erythrocyte; gdT, γδ T cell; HSPC, human stem and progenitor cell; TCM, central memory T cell; TEM, effector memory T cell; T_{reg}, regulatory T cell.

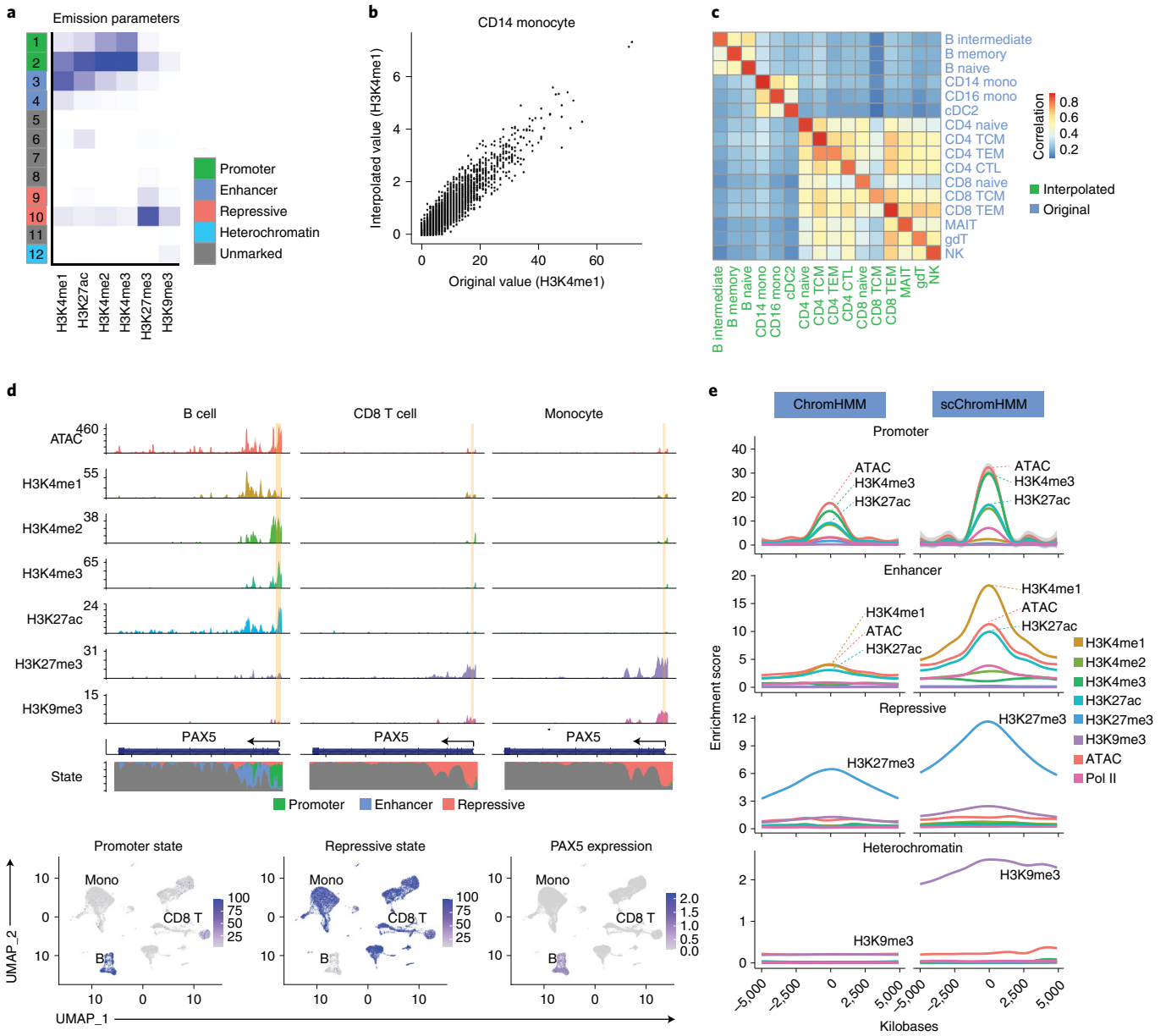


Fig. 3 | scChromHMM annotates chromatin states at single-cell resolution. a, Chromatin states returned by ChromHMM, which was run on 25 pseudobulk tracks for six histone marks. States are broadly grouped into five categories. **b**, Correlation comparing cell-type-specific pseudobulk profiles of H3K4me1 in CD14 monocytes generated from the original experiment or the interpolated values. Each point corresponds to a 200-bp genomic window (Methods) (**c**), For each cell type, the interpolated and original profiles are highly correlated and clustered together. **d**, scChromHMM outputs at the PAX5 locus. Top: Pseudobulk profiles for six chromatin marks in three cell types. Yellow bar highlights a 200-bp genomic window near the TSS. Bottom: scChromHMM posterior probabilities representing the annotation for the highlighted window in each cell. The region is uniformly annotated with a promoter state in B cells, where PAX5 is transcriptionally active, and as a repressive state in other cell types. **e**, Metaplots exhibiting the enrichment of chromatin accessibility and histone modifications at functional regions identified by ChromHMM (left) and scChromHMM (right) in CD14 monocytes. K, kilobase.

a form of data 'denoising' by averaging histone modification signals across cells in similar biological states and therefore alleviates the sparsity limitations associated with scCUT&Tag measurements. We emphasize that the acquisition of protein measurements is essential for this strategy, as the protein modality enables the accurate identification of anchors across diverse datasets.

Having acquired interpolated chromatin profiles for six histone modifications at single-cell resolution, we next aimed to annotate the chromatin state of each 200-bp genomic window in each cell. To achieve this, we ran the forward–backward algorithm[12,36] individually on each cell (Methods) using our interpolated profiles for six histone modifications and our previously calculated set of emission and transition probabilities as input. We note that data binarization is a requirement for this procedure and found that the frequency of our interpolated profiles across genomic windows exhibited clear evidence of bimodality, facilitating this process (Supplementary Fig. 6a). The output of scChromHMM represents, for each genomic window in each cell, the posterior probability distribution across 12 chromatin states. For example, in Fig. 3d, we visualize the posterior probabilities (promoter) of a genomic window located at the PAX5 transcriptional start site (TSS), at single-cell resolution.

Regions classified by scChromHMM as active promoters were enriched near the TSS of actively transcribed genes, active

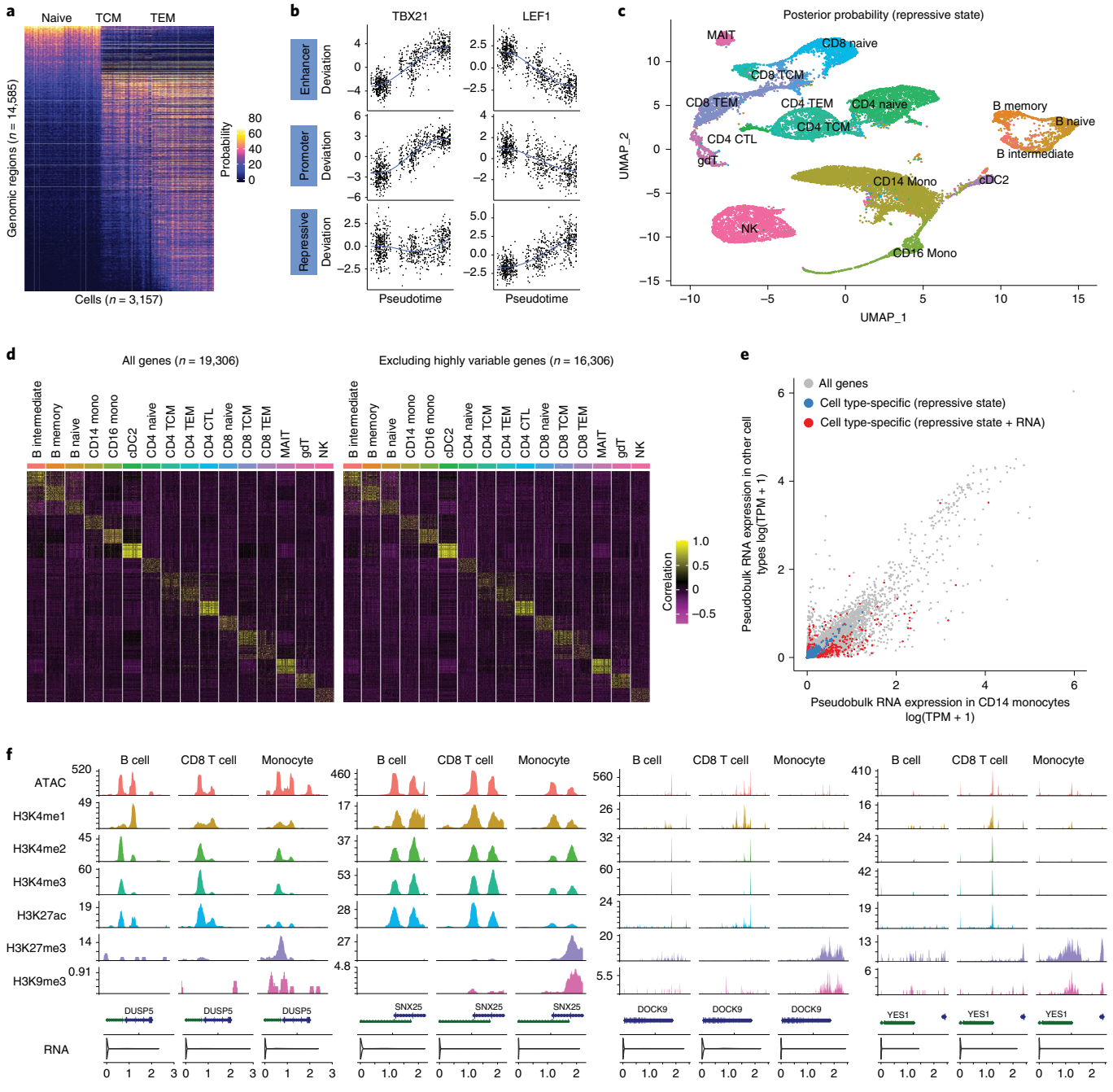


Fig. 4 | Extensive heterogeneity in repressive chromatin encodes cellular identity. a, Remodeling of repressive chromatin during CD8 T cell maturation. Heatmap shows the posterior probabilities (repressive state) in single cells for 14,585 genomic loci, as returned by scChromHMM. Cells are ordered by their progression along pseudotime (Fig. 2c). **b**, chromVAR deviation scores for the TBX21 and LEF1 motifs in single cells, ordered by their progression along pseudotime. We used the scChromHMM-derived posterior probabilities as input to chromVAR instead of chromatin accessibility levels. **c**, Unsupervised analysis of scChromHMM-derived probabilities (repressive state) separates granular cell types. **d**, Single-cell correlation matrix based on repressive chromatin at TSSs (Methods) when using all TSSs (left heatmap) or after excluding the top 3,000 transcriptionally variable genes (right heatmap). In each case, the observed correlation structure is fully consistent with cell type labels, suggesting that there is extensive heterogeneity in repressive chromatin even for genes that do not vary transcriptionally. **e**, Scatter plot showing average gene expression levels for all genes in CD14 monocytes (*x* axis) and other cell types (*y* axis). Colored points represent 1,597 loci where we detect changes in repressive chromatin for monocytes (Methods). Blue points represent 1,340 loci where we do not detect an accompanying transcriptional change. Red points represent 257 genes where we detect a transcriptional shift. TPM, transcripts per kilobase million. **f**, Four representative examples of individual genes shown as blue points in **e**.

enhancer regions were distal to the TSS but enriched for accessible chromatin, and heterochromatic regions largely overlapped (76.3%) with a set of repetitive elements annotated by RepeatMasker[39] (Fig. 3e and Methods). We also found that putatively functional regions identified by scChromHMM exhibited increased accuracy and robustness compared to regions identified by ChromHMM run on pseudobulk profiles. Regions classified by scChromHMM consistently exhibited stronger enrichment for key histone

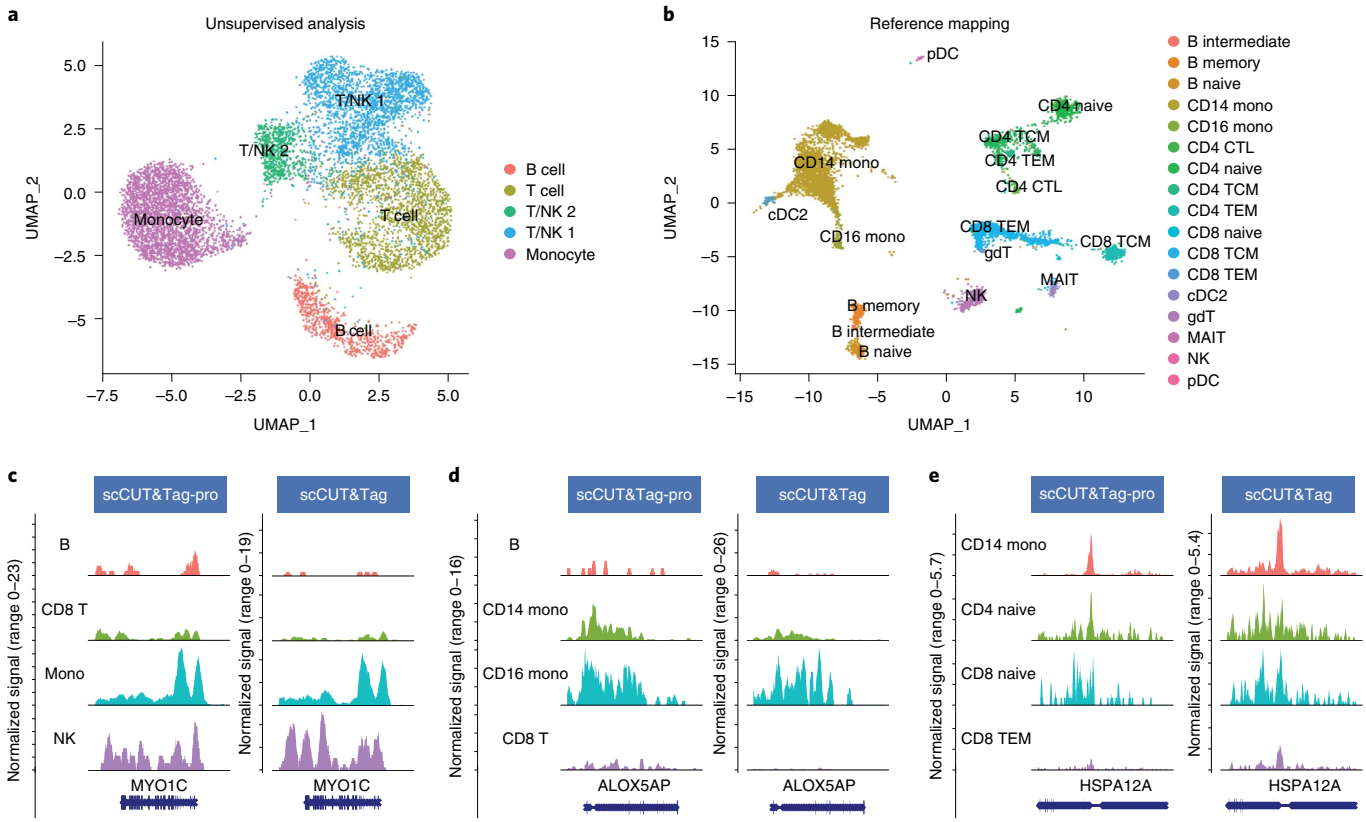


Fig. 5 | Supervised mapping of scCUT&Tag datasets. a, UMAP visualization of 8,362 H3K27me3 scCUT&Tag profiles of human PBMCs from Wu et al.[19] based on an unsupervised analysis and clustering. **b**, Same cells as in **a**, but after mapping to the multimodal reference defined in this paper. Cells are colored by their reference-derived level 2 annotations. **c**–**e**, Coverage plots showing the cell-type-specific binding patterns of H3K27me3 at three loci, MYO1C (**c**), ALOX5AP (**d**) and HSPA12A (**e**). Plots are shown for our dataset (reference), as well as the scCUT&Tag profiles from the query dataset (query, Wu et al.[19]). Cells in the query dataset are grouped by their predicted labels. We observe highly concordant patterns across datasets for all loci, supporting the accuracy of our predictions. Four representative cell types are shown at each locus. TEM, T effector memory.

modifications (Fig. 3e) and RNA polymerase II (Supplementary Fig. 6b) compared to active promoters identified by ChromHMM. We conclude that scChromHMM enables chromatin state predictions at single-cell resolution and also improves the accuracy of state prediction.

Heterogeneity in repressive chromatin suggests regulatory priming. We used the output of scChromHMM to explore the dynamics of chromatin state changes across a trajectory of CD8 T cell maturation from naive to effector states. For example, we identified 14,585 genomic windows that acquired or lost a repressive chromatin state (Methods), indicating that extensive remodeling of repressive chromatin is associated with this cellular process, particularly at the sharp transition from naive to memory cells (Fig. 4a). To explore motifs that were associated with these transitions, we modified the chromVAR algorithm[40] to identify motifs whose presence was correlated with an increased posterior probability across the learned chromatin states (Methods).

Our analyses revealed that genomic windows containing the motif bound by the transcription factor TBX21 (T-bet), which is a crucial regulator of T cell responses[41,42], exhibited highly dynamic chromatin states during T cell maturation (Fig. 4b). In particular, as cells transition from naive to memory and effector states, genomic windows containing TBX21 motifs acquired an increased probability of entering an active promoter or enhancer state. Conversely, genomic windows containing motifs bound by the critical T cell regulator LEF1 (ref. [43]) tended to exist in active chromatin states at the onset of the trajectory, but the same windows transitioned to acquire a repressive state as cells progressed through the trajectory (Fig. 4b). More generally, when aggregating cells into granular subgroups, we observed substantial cellular heterogeneity and cell type specificity for motif activity scores computed based on scChromHMM-determined probabilities (Supplementary Fig. 6c). This analysis nominated a suite of regulators for each cell type that were also differentially expressed at the transcriptional level (Supplementary Fig. 6c). Taken together, our results demonstrate the potential for integrated analyses of scCUT&Tag data to identify DNA sequence motifs that correlate with diverse chromatin states.

We next asked if cellular heterogeneity in repressive chromatin state was fully reflected in the cellular transcriptome. We first considered the scChromHMM-derived probabilities for the repressive state at genomic windows overlapping TSSs. We found that granular cellular states could be separated based on unsupervised clustering of these measurements (Methods), indicating that cellular identity is encoded not only in transcriptional output or immunophenotypes but also in genome-wide profiles of repressive chromatin state (Fig. 4c). To our surprise, we found that granular cellular identity remained encoded in repressive chromatin probabilities at TSSs, even after excluding the 3,000 most variable genes in the transcriptome (Fig. 4d). This suggested that even for genes whose expression did not vary across cell types, there may be substantial heterogeneity in chromatin state.

We therefore considered the set of loci where we observed cell-type-specific changes in repressive chromatin state (Methods)

and calculated their overlap with transcriptional changes as well. For example, we observed 1,597 TSS loci whose posterior probability (repressive chromatin state) and H3K27me3 signal was enriched or depleted in CD14 monocytes. Within the associated genes for these loci, we detected monocyte-specific transcriptional shifts in only 257 (16.1%) (Fig. 4e,f). When we did observe both transcriptional and chromatin-based shifts, they were highly concordant (i.e., the acquisition of a repressive state was associated with a decrease in gene expression). However, when we identified heterogeneity in repressive chromatin state in the absence of transcriptional variation, gene expression was either absent or at very low levels in all cell types (Fig. 4e). When considering their pseudobulk expression, the median expression of the set of 1,340 genes in PBMCs was below 0.5 transcripts per million. These genes are not informative when determining cellular state from scRNA-seq data, yet their repressive chromatin landscape exhibits clear patterns of cell type specificity (Fig. 4f).

We note that some genes exhibiting this pattern may in fact be differentially expressed across cell types but cannot be accurately quantified by scRNA-seq. To address this, we analyzed a publicly available transcriptome atlas dataset[44], with high-quality bulk total RNA-seq profiles for sorted PBMC subsets. Of the 1,340 genes for which we detected monocyte-specific changes in repressive chromatin, but not transcription, 1,081 (81%) failed to show evidence of differential expression in bulk data as well (DESeq2 adjusted $P > 0.01$), demonstrating that our findings are not driven by artifacts driven by scRNA-seq.

Gene ontology analysis[45] of these gene sets was enriched for DNA-binding transcription factors and transmembrane and channel proteins whose activity may regulate the activation of downstream signaling pathways and immune responses (Supplementary Fig. 6d). In a meta-analysis of these TSSs, monocytes exhibited enrichment for H3K27me3, weak enrichment of H3K4me3 and H3K4me2 and the absence of H3K27ac (Supplementary Fig. 6e). We therefore hypothesize that these chromatin shifts may represent a form of regulatory 'poising' that is not yet reflected in the current transcriptional output of each cell but may be relevant to its future behavior or potential. Our findings are consistent with previous studies of the epigenetic landscape in human T cells, which have shown that the promoters of many silent genes were enriched for active histone modifications and could be rapidly induced in response to environmental stimuli[46]. Moreover, studies in differentiating T cells have found that distinct and lineage-specific chromatin states can be present even for nonexpressed genes[47–49].

Projecting scCUT&Tag datasets onto a multimodal atlas. Our analyses demonstrated the utility of projecting data from diverse molecular modalities into a harmonized manifold, creating a multimodal atlas that spans CITE-seq, ASAP-seq and scCUT&Tag-pro experiments. However, it is not always possible to measure cell surface protein levels when performing scCUT&Tag, particularly when performing analysis of single nuclei. We therefore considered how to project unimodal scCUT&Tag profiles into our reference dataset.

We have recently released a framework, Azimuth[30], which addresses a similar problem of projecting scRNA-seq datasets onto a CITE-seq-defined human PBMC reference. In Azimuth, we use a semi-supervised dimensionality reduction approach, supervised principal component analysis (PCA)[50], which identifies the best transcriptomically defined vectors that separate CITE-seq defined cell types. We pursued a similar strategy to map a recently published H3K27me3 scCUT&Tag dataset of human PBMC nuclei[19] onto our multimodal reference atlas. Using our scCUT&Tag-pro dataset as a reference, we calculated a modified LSI (Methods), which transforms the histone modification profiles into a low-dimensional space that retains separation of our multimodally defined cell types. We used this transformation to identify anchors and 'map' the query dataset into our reference dataset (Methods).

We observed that by performing reference mapping, we could substantially improve the interpretation of the query dataset, particularly the ability to resolve granular cell types. For example, unsupervised clustering of the query dataset[19] was consistent with the previously published analysis and revealed broad clusters of immune subsets, including monocyte, B cell and T cell/NK cell subgroups (Fig. 5a). After reference mapping, monocytes could be segregated into CD14 and CD16$^+$ subsets, T cell/NK cell subclusters could be further resolved into CD4 T, CD8 T and NK cell groups with additional heterogeneity for naive and effector states and B cells separated into different developmental stages (Fig. 5b). We found that we could classify (prediction score >0.5) 88% of cells at level 1 resolution and 73% of cells at level 2 resolution. We find high concordance when comparing patterns of cell-type-specific H3K27me3 binding compared in both the reference and query datasets (Fig. 5c–e), verifying the accuracy of our predictions. We note that although we transfer discrete annotations in this example, it is also possible to transfer continuous data, which would enable the prediction of one histone modification based on the measurement of another. Therefore, we propose that our multimodal atlas represents a broad scaffold for the circulating human immune system that can be used to map datasets from a wide variety of cutting-edge single-cell technologies.

Discussion

One of the key goals of multiomic single-cell studies is to explore the relationships across different molecular modalities and how changes in one modality affect variation in another. Here, we pursue an integrative analysis where nine different molecular modalities are each measured in separate experiments. We subsequently integrate these data and derive what we colloquially refer to as single-cell 'megaomic' profiles, each representing an individual cell containing measured or interpolated measurements for many modalities. In addition, we introduce scChromHMM, which enables the exploration of heterogeneity in chromatin state across both discrete cell types and continuous trajectories.

Substantial recent progress has been made in simultaneous profiling of multiple modalities in single cells[22,23,51,52], but currently, it is not feasible to simultaneously profile the genome-wide binding patterns of six different histone modifications in the same cell, particularly when there is a partial overlap in their localization. Additionally, future extensions of CUT&Tag enabling simultaneous measurements of RNA and protein levels are likely to result in decreased per-modality data quality, as has been previously described[28,29]. Our approach therefore represents a feasible and broadly applicable alternative, and we demonstrate that our megaomic profiles are highly quantitative, sensitive and robust. However, interpolated modality predictions cannot capture stochastic technical or biological variation or be used to detect associations between multiple histone modifications within the same cell.

The successful integration of CITE-seq, ASAP-seq and scCUT&Tag-pro datasets originates from a shared panel of cell surface protein levels that are collected in each experiment. We emphasize that cell surface protein levels represent one of multiple potential modalities that can be used for megaomics integration. For example, two recent pioneering approaches (Paired-Tag[22] and CoTECH[23]) introduced combinatorial indexing-based approaches for simultaneous scCUT&Tag and scRNA-seq in single cells, enabling the integration of additional multiomic technologies (i.e., Paired-Seq) via transcriptomic measurements. Our scCUT&Tag-pro technology represents a complementary solution that is compatible with the 10x Genomics Chromium system, and will be particularly valuable for profiling immune cell types that are well defined by their surface protein landscape.

We found substantial heterogeneity in chromatin state even at genes that did not exhibit transcriptional variation. In particular, we observed cell-type-specific repression, as demarcated by the presence of H3K27me3, even for genes that were expressed at similarly low levels across all cell types. Although it is possible that this heterogeneity in repressive chromatin state has minimal functional consequence, cellular variation in chromatin state may also underlie future cell type-specific transcriptional responses. In addition to previous findings in human immune cells[46–49], Mellis et al. found that cell-type specificity was encoded not only in steady-state transcriptional output but also in future responses to environmental perturbations[53]. Similarly, Shim et al. found that genes with heterogeneous H3K27me3 landscapes across distinct organ systems were enriched for key developmental, regulatory and morphological functions[54]. We propose that scCUT&Tag-pro represents a sensitive technique for identifying loci exhibiting heterogeneous repressive chromatin within an organ system as well. Future experiments profiling chromatin state in the bone marrow will help to elucidate the developmental dynamics that establish this heterogeneity.

Moving forward, we expect that integrated analyses across multiple modalities will not only help to understand the molecular state of single cells but also reveal new insights on fundamental questions in gene regulation. For example, the activation of gene expression is correlated with changes in chromatin accessibility, the acquisition and loss of histone modifications, the binding of transcription factors and RNA polymerase and the formation of new DNA contacts. However, the regulatory relationships and temporal ordering between these steps remains poorly understood. Motivated by the ability of RNA velocity analysis to use known temporal relationships between unspliced and spliced reads to infer the direction of developmental trajectories[55], we envision that single-cell multiomic analysis will help to order the coordinated molecular events that collectively establish cellular heterogeneity.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01250-0.

Received: 14 September 2021; Accepted: 7 February 2022;
Published online: 24 March 2022

References

1. Tang, F. et al. mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
2. Jaitin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
3. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
4. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
5. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
6. Shalek, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
7. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
8. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
9. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
10. Hoffman, M. M. et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).
11. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).
12. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).
13. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
14. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
15. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).
16. Kaya-Okur, H. S. et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
17. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
18. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.* **39**, 825–835 (2021).
19. Wu, S. J. et al. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. *Nat. Biotechnol.* **39**, 819–824 (2021).
20. Wang, Q. et al. CoBATCH for high-throughput single-cell epigenomic profiling. *Mol. Cell* **76**, 206–216 (2019).
21. Carter, B. et al. Mapping histone modifications in low cell number and single cells using antibody-guided chromatin tagmentation (ACT-seq). *Nat. Commun.* **10**, 3747 (2019).
22. Zhu, C. et al. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat. Methods* **18**, 283–292 (2021).
23. Xiong, H. et al. Single-cell joint detection of chromatin occupancy and transcriptome enables higher-dimensional epigenomic reconstructions. *Nat. Methods* **18**, 652–660 (2021).
24. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* **12**, 7–18 (2011).
25. Suganuma, T. & Workman, J. L. Signals and combinatorial functions of histone modifications. *Annu. Rev. Biochem.* **80**, 473–499 (2011).
26. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
27. Peterson, V. M. et al. Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **35**, 936–939 (2017).
28. Mimitou, P. et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* **39**, 1246–1258 (2021).
29. Swanson, E. et al. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *Elife* **10**, e63632 (2021).
30. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
31. Nathan, A. et al. Multimodally profiling memory T cells from a tuberculosis cohort identifies cell state associations with demographics, environment and disease. *Nat. Immunol.* **22**, 781–793 (2021).
32. Pombo Antunes, A. R. et al. Single-cell profiling of myeloid cells in glioblastoma across species and disease stage reveals macrophage competition and specialization. *Nat. Neurosci.* **24**, 595–610 (2021).
33. Lareau, C. A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).
34. Stoeckius, M. et al. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
35. Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
36. Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge Univ. Press, 1998).
37. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022).
38. Stuart, T. et al. Comprehensive inteÿgration of single-cell data. *Cell* **177**, 1888–1902 (2019).
39. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0 (ISB, 2013); http://www.repeatmasker.org
40. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
41. Szabo, S. J. et al. A novel transcription factor, T-bet, directs Th1 lineage commitment. *Cell* **100**, 655–669 (2000).
42. Kallies, A. & Good-Jacobson, K. L. Transcription factor T-bet orchestrates lineage development and function in the immune system. *Trends Immunol.* **38**, 287–297 (2017).

43. Zhao, X., Shan, Q. & Xue, H.-H. TCF1 in T cell immunity: a broadened frontier. *Nat. Rev. Immunol.* https://doi.org/10.1038/s41577-021-00563-6 (2021).
44. Lorenzi, L. et al. The RNA Atlas expands the catalog of human non-coding RNAs. *Nat. Biotechnol.* **39**, 1453–1465 (2021).
45. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
46. Roh, T.-Y., Cuddapah, S., Cui, K. & Zhao, K. The genomic landscape of histone modifications in human T cells. *Proc. Natl Acad. Sci. USA* **103**, 15782–15787 (2006).
47. Wei, G. et al. Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity* **30**, 155–167 (2009).
48. Schoenborn, J. R. et al. Comprehensive epigenetic profiling identifies multiple distal regulatory elements directing transcription of the gene encoding interferon-gamma. *Nat. Immunol.* **8**, 732–742 (2007).
49. Cui, K. et al. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* **4**, 80–93 (2009).
50. Barshan, E., Ghodsi, A., Azimifar, Z. & Zolghadri Jahromi, M. Supervised principal component analysis: visualization, classification and regression on subspaces and submanifolds. *Pattern Recognit.* **44**, 1357–1371 (2011).
51. Gopalan, S. et al. Simultaneous profiling of multiple chromatin proteins in the same cells. *Mol. Cell* **81**, 4736–4746 (2021).
52. Meers, M. P., Janssens, D. H. & Henikoff, S. Multifactorial chromatin regulatory landscapes at single cell resolution. Preprint at *bioRxiv* https://doi.org/10.1101/2021.07.08.451691 (2021).
53. Mellis, I. A. et al. Responsiveness to perturbations is a hallmark of transcription factors that maintain cell identity in vitro.*Cell Syst.* **12**, 885–899 (2021).
54. Shim, W. J. et al. Conserved epigenetic regulatory logic infers genes governing cell identity. *Cell Syst.* **11**, 625–639 (2020).
55. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
57. Stuart, A. et al. Multimodal single-cell chromatin analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).

Methods

Protocols. A step-by-step online protocol for scCUT&Tag-pro is available on the protocols.io platform (https://www.protocols.io/private/57E45A034D5311ECA43B0A58A9FEAC02).

PBMC acquisition and processing. Cryopreserved healthy donor PBMCs were purchased from AllCells. After thawing into DMEM with 10% FBS, the cells were spun down at 4 °C for 5 min at 400 *g* and washed twice with PBS with 2% BSA. After centrifugation, the cell pellet is resuspended in staining buffer (2% BSA and 0.01% Tween in PBS).

scCUT&Tag antibodies. The antibodies used were H3K4me1 (1:100, Abcam, ab8895), H3K4me2 (1:100, Abcam, ab32356), H3K4me3 (1:100, Abcam, ab213224), H3K27ac (1:100, Abcam, ab177178), H3K27me3 (1:100, Cell Signaling Technology, 9733), H3K9me3 (1:100, Abcam, ab8898), Phospho-Rpb1 CTD (Ser2/Ser5) (1:50, Cell Signaling, 13546) and guinea pig anti-rabbit (1:100, Novus Biologicals, NBP1-72763). TotalSeq-A-conjugated antibodies and panels were obtained from BioLegend (399907; Supplementary Table 2 lists the antibodies, clones and barcodes used).

scCUT&Tag-pro experimental workflow. *Surface protein antibody staining.* Five million thawed PBMCs were resuspended in 200 µl staining buffer (2% BSA and 0.01% Tween in PBS) and incubated for 15 min with 10 µl Fc receptor block (BioLegend, TruStain FcX) on ice. The samples were evenly distributed into five tubes, and 2 µl hashing antibody was added separately. Cells were then washed three times with 1 ml staining buffer and pooled together. Two million PBMCs were used for each histone modification. After cell hashing, the panel of oligo-conjugated antibodies (one test is sufficient for two million cells) was added to the cells to incubate for 30 min on ice. After staining, cells were washed three times with 1 ml staining buffer and resuspended in 100 µl staining buffer. Subsequently, 4 µl Fab fragment goat anti-mouse IgG (Jackson ImmunoResearch, 115-007-003) was added for incubation on ice for 15 min. The cells were then washed three times with 1 ml staining buffer. After the final wash, cells were resuspended in 200 µl PBS and ready for fixation.

Fixation and permeabilization. A total of 1.25 µl 16% methanol-free formaldehyde (Thermo Fisher Scientific, PI28906) was added for fixation (final concentration, 0.1%) at room temperature for 5 min. The cross-linking reaction was stopped by addition of 12 µl 1.25 M glycine solution. Subsequently, cells were washed twice with PBS. The permeabilization was performed by adding isotonic lysis buffer (20 mM Tris-HCl, pH 7.4, 150 mM NaCl, 3 mM $MgCl_2$, 0.1% NP-40, 0.1% Tween-20, 1% BSA and 1× protease inhibitors) on ice for 7 min. Subsequently, 1 ml cold wash buffer (20 mM HEPES, pH 7.6, 150 mM NaCl, 0.5 mM spermidine and 1× protease inhibitors) was added, and cells were centrifuged at 800 *g* for 5 min at 4 °C.

Tagmentation. Permeabilized cells were directly resuspended with 150 µl antibody buffer (20 mM HEPES, pH 7.6, 150 mM NaCl, 2 mM EDTA, 0.5 mM spermidine, 1% BSA and 1× protease inhibitors) with primary antibody (e.g., anti-H3K4me1) and incubated overnight on a rotator at 4 °C. The next day, cells were washed once with 150 µl wash buffer and centrifuged for 5 min at 800 *g*. After removing the supernatant, the cells were resuspended in 150 µl wash buffer with secondary antibody and incubated for 1 h at room temperature on a rotator. The cells were washed twice with 150 µl wash buffer to remove excess remaining antibodies. The cells were then resuspended in 150 µl high-salt wash buffer (20 mM HEPES, pH 7.6, 300 mM NaCl, 0.5 mM spermidine and 1× protease inhibitors) with 7.5 µl pAG-Tn5 (EpiCypher, 15-1017) and incubated for 1 h on rotator at room temperature. The cells were then washed twice with high-salt wash buffer and resuspended in 100 µl tagmentation buffer (20 mM HEPES, pH 7.6, 300 mM NaCl, 0.5 mM spermidine, 10 mM $MgCl_2$ and 1× protease inhibitors). The samples were placed in a PCR machine and incubated for 1 h at 37 °C. When tagmentation was done, the reaction was stopped by adding 4 µl 0.5 M EDTA. Tagmentation steps were performed in 0.2-ml tubes to minimize cell loss.

Single-cell encapsulation, PCR and library construction. After tagmentation, cells were centrifuged for 5 min at 1,000 *g*, and the supernatant was discarded. Cells were resuspended with 30 µl 1× Diluted Nuclei Buffer (10x Genomics), counted and diluted to a concentration based on the targeted cell number. Seven microliters of ATAC buffer B was added to 8 µl of cells, pipetting several times to mix. All remaining steps were performed according to the 10x Chromium single-cell ATAC protocol and the library construction method was adapted from ASAP-seq[28]. Briefly, 0.5 µl 1 µM bridge oligo A (TCGTCGGCAGCGTCAGATG TGTATAAGAGACAGNNNNNNNNVTTTTTTTTTTTTTTTTTTTTTTT TTTTTTT/3InvdT/) was added to the barcoding mix. Linear amplification was performing using the following PCR program: 40 °C for 5 min, 72 °C for 5 min, 98 °C for 30 s; 12 cycles of 98 °C for 10 s, 59 °C for 30 s and 72 °C for 1 min; ending with hold at 15 °C. The remaining steps were performed according to the 10x scATAC-seq protocol (v1.1), with the following additional modifications.

For ADTs, during silane bead elution (step 3.1o), beads were eluted in 43.5 µl Elution Solution I. The extra 3 µl was used for the surface protein tag library. During SPRI cleanup (step 3.2d), the supernatant was saved, and the short DNA derived from antibody oligos was purified with 2.0× SPRI beads. The eluted DNA was combined with the 3 µl left aside after silane purification to be used as input for protein tag amplification. PCR reactions were set up to generate the protein tag library with KAPA Hifi Master Mix (P5 and RPI-x primers) using the following program: 95 °C for 3 min; 14–16 cycles of 95 °C for 20 s, 60 °C for 30 s and 72 °C for 20 s; followed by 72 °C for 5 min and ending with hold at 4 °C.

The P5 primer sequence was 5′-AATGATACGGCGACCACCGAGATCTACAC-3′, and the RPI-x primer sequence was 5′-CAAGCAGAAGACGGCATACGAGATxxxxxxxxGTGACTG GAGTTCCTTGGCACCCGAGAATTCCA-3′.

Sequencing. The final libraries (CUT&Tag library, ADT library and HTO library) were sequenced on NextSeq 550 with the following recipe: i5:16 bp, i7: 8 bp, read1: 34 bp, read2: 34 bp.

scCUT&Tag-pro data preprocessing. *ADTs and cell hashing libraries.* We used salmon alevin 1.4.0 (ref. [56]) for the alignment and quantification of single-cell ADTs and hashtag oligos (HTOs) using a custom index that is constructed only from known ADT and HTO barcode sequences. We applied a threshold of ten protein and ten hashing antibody counts. Cells that passed these thresholds were used as input to the HTODemux() function in Seurat[30] for demultiplexing and doublet removal. Antibody data were normalized using the centered log ratio transformation. To visualize cells on the basis of their ADT profiles, we performed PCA using all features as input and retained the first 25 principal components to construct a nearest neighbor graph and perform UMAP visualization.

pAG-Tn5 tagmented libraries (i.e., scCUT&Tag). We used the cellranger-atac v1.2 pipeline with default settings to align reads to the hg38 human reference annotation and generate fragment files. We filtered the fragment file to keep the cellular barcodes that were present in the filtered ADT and HTO libraries and also contained at least 100 fragments. We use the AggregateTiles function from the Signac package[57] to quantify the fragment coverage based on 5,000-bp genomic windows, remove windows with less than five counts and then merge adjacent windows. We next performed modified term frequency–inverse document frequency (TD-IDF) normalization[57], which corrects for differences in cellular sequencing depth and weights each window based on its average fragment abundance across all cells. We next ran singular value decomposition (SVD) on the TF-IDF matrix to generate the low-dimensional representation of the data. We used dimensions 2:30 for UMAP visualization, as the first dimension is highly correlated with sequencing depth.

We used the Seurat v4 WNN framework[30] to generate a multimodal representation of the scCUT&Tag-pro datasets. We used the FindMultiModalNeighbors function to generate a WNN graph using dimensions 1:25 (ADT modality) and 2:30 (scCUT&Tag modality).

Reference mapping. We integrated each of our scCUT&Tag-pro datasets with a previously published reference CITE-seq dataset[30]. In previous work, we applied supervised PCA (sPCA) to identify a linear projection of the scRNA-seq measurements in the CITE-seq dataset that maximally retained the information captured in both modalities. This transformation can be applied to new scRNA-seq query datasets to identify anchors with the CITE-seq reference. These anchors were used for annotation and visualization of the query cells. Our framework for query-to-reference mapping was implemented as part of our Azimuth tool, which is accessible at azimuth.hubmapconsortium.org and fully described in our previous paper[30].

Here, we used a similar strategy but performed sPCA on the protein measurements from the CITE-seq dataset. As input to sPCA, we used all 128 protein features that were measured both in this study and in the reference dataset. This procedure identified a linear transformation of cell surface protein measurements that can be applied to scCUT&Tag-pro measurements. We used the FindTransferAnchors anchors (reduction = 'spca', dims = 1:30) in Seurat v4 to identify anchors between reference and query datasets and the MapQuery function (default parameters) to annotate query cells at multiple levels of resolution and visualize these cells in a 2D UMAP. After all datasets were mapped, we subsetted all CD8 T cells (which encompass memory, naive and effector states). We used monocle 3 with default parameters[58] to identify an integrated developmental trajectory that orders cells from all experiments.

Interpolating multiple histone modification profiles in single cells. To perform an integrated analysis of chromatin state at single-cell resolution, we first generated a set of single-cell profiles that comprised interpolated, genome-wide and quantitative measurements of six histone modifications. In previous work[38], we demonstrated how to use anchors between reference and query datasets to transfer not only discrete information (i.e., cell-type annotations) but also quantitative measurements. For example, we demonstrated how to accurately interpolate cell surface protein measurements for a scRNA-seq dataset of 274,599 human bone marrow mononuclear cells generated by the Human Cell Atlas based on a CITE-seq dataset of 35,543 cells from the same system.

We applied a similar workflow for this study (schematized in Fig. 2a). First, we considered a CITE-seq dataset[30], where each cell contained transcriptomic measurements alongside quantifications for a panel of 228 surface proteins. Although the original dataset contained 161,764 cells, interpolating genome-wide profiles for multiple modalities at this scale created memory and storage constraints, and we therefore selected a random subset of 20,000 cells for downstream analysis. In Supplementary Fig. 5c, we demonstrate that we obtained highly concordant results when repeating the procedure on a second random downsample of 20,000 cells.

We then found anchors between each scCUT&Tag-pro dataset and the 20,000 CITE-seq cells. Specifically, the 'anchors' in this procedure represent pairwise correspondences of cells: a CITE-seq cell that lacks the histone modification profile of interest and an scCUT&Tag-pro cell in a matched biological state, but where the histone modification was directly measured. Anchors were determined using the FindTransferAnchors function, and we used the same protein sPCA transformation described in 'Reference mapping'. We used these anchors to interpolate (i.e., 'impute') the number of fragments for each histone modification, in each 200-bp genomic window, at single-cell resolution (see Stuart et al.[38], 'Feature Imputation'). For example, in each of the 20,000 CITE-seq cells, the interpolated levels of H3K4me1 represent a weighted average of the anchor-forming cells in the H3K4me1 scCUT&Tag-pro. We repeated this procedure for each of the six modifications, as well as the ASAP-seq dataset[28]. Therefore, each of the 20,000 cells contained information for nine modalities, encompassing both the original measurements (RNA + protein) and seven interpolated modalities representing histone modifications and chromatin accessibility.

Based on our pseudobulk saturation analysis (Fig. 1e), we used a weighted average of 500 anchors when interpolating values for each cell. This weighting prioritizes anchors that are most similar to the query cell, enabling us to accurately interpolate histone modification profiles even for cell types present at moderate abundance (Fig. 3c and Supplementary Fig. 5a). However, for rare cell types, there is a possibility that an insufficient number of anchors may reduce the accuracy of these predictions. We therefore removed cell types that were present at <2.5% frequency in our reference dataset from further analysis.

In Fig. 3b, we plot the correlation between interpolated and original values for the H3K4me1 histone modification. Each point represents a 200-bp genomic window. To avoid overplotting, we only plotted windows on chr22 ($n = 254{,}093$).

Assigning chromatin states at single-cell resolution. To integrate information from multiple histone modifications, we used the multivariate HMM introduced in ChromHMM[12]. For clarity, we reintroduce the key components of the model here. The model assumes that each 200-bp genomic window ('genomic interval') can be represented based on one of K hidden states, where each state is defined by the combinatorial presence or absence of multiple histone modifications. The full likelihood of the model is:

$$(v|a,b,p) = \prod_{c\in C}\sum_{s_c\in S_c} a_{s_{c_1}}\left(\prod_{t=2}^{T_C} b_{s_{c_{t-1}},s_{c_t}}\right)\prod_{t=1}^{T_C}\prod_{m=1}^{M} p_{s_{c_t},m}^{v_{c_t,m}}\left(1-p_{s_{c_t},m}\right)^{(1-v_{c_t,m})},$$

where v represents the observed data, such that $v_{c_t,m}$ is 1 if mark m is present at genomic interval c_t and 0 otherwise. $p_{k,m}$ represents the probability that mark m is present at a genomic interval in state k ('emission probability'), $b_{i,j}$ represents the probability of the Markov chain transitioning from state i to state k ('transition probability') and $s_{c,t}$ represents the 'hidden state' associated with each at genomic interval c_t.

The first step of the procedure is to learn a set of possible chromatin states and their associated emission and transition probabilities. We used ChromHMM's variant of the Baum–Welch algorithm[12] as implemented in the LearnModel command, setting the number of states to 12 and the maximum number of iterations to 1,000. As input to ChromHMM[17], we provided the histone modification profiles for all cell types, where cell type identity was determined after reference mapping (level 2 resolution). The 12 learned states and associated emission probabilities are shown in Fig. 3a.

After learning a set of chromatin states, we next inferred the path of hidden states (hidden state sequence) through each chromosome at single-cell resolution. To do so, we extended the ChromHMM method to calculate the posterior distribution of state assignments for each 200-bp genomic interval in each cell. We refer to the following procedure as scChromHMM:

(1) The input to scChromHMM is the set of interpolated histone modifications in single cells, as described above.
(2) As ChromHMM models multiple chromatin marks as an independent set of Bernoulli random variables, we must binarize the interpolated values present in each cell. As shown in Supplementary Fig. 6a, the distribution of abundances for each histone modification exhibited clear bimodality. We assigned binarization thresholds via manual inspection (red dashed line), which were congruent with the results of Hartigan's dip test. Binarization thresholds were set independently for each mark but kept constant across single cells. After thresholding, each genomic interval in each cell is represented by the presence or absence of six histone modifications.
(3) We ran the forward–backward algorithm (a dynamic programming algorithm to learn the posterior probabilities of hidden state variables in an HMM[36]) on each cell independently. The output of this procedure is a posterior probability distribution over all 12 states for each genomic interval in each cell.

Comparison of scChromHMM and ChromHMM results. In addition to scChromHMM, we also annotated genomic intervals using the standard ChromHMM workflow[17]. Although this approach does not return state predictions in individual cells, it does return posterior probability distributions for each cell type (level 2 resolution). We therefore compared the predictions from ChromHMM[17] with the predictions from scChromHMM after grouping by cell type. In Fig. 3e, we perform a meta-analysis across all regions assigned to promoter, enhancer, repressor and heterochromatic states and explore the level of enrichment for each modification in a 10-kb window centered on these regions. We combined the posterior probabilities for functionally related states (promoters, states 1 and 2; enhancers, states 3 and 4; and repressor, states 9 and 10).

We compare results on CD14 monocytes, as this was the most abundant cell type in our dataset. For ChromHMM[17], we selected all regions with >75% probability posterior probability in CD14 monocytes for meta-analysis, and for scChromHMM, we selected all regions where the majority of CD14 monocytes had >75% posterior probability. In Supplementary Fig. 6b, we also compare average polymerase II fragment abundance as measured by scCUT&Tag-pro on promoter regions as identified by either ChromHMM[17] or scChromHMM.

Identifying heterogeneity in chromatin state across a continuous trajectory. In previous work[38], we used the Moran's I statistic measure of spatial autocorrelation to identify genes whose expression varied as a function of spatial location. Here, we used a similar approach to identify 14,585 genomic intervals whose posterior probability (repressive state) varies across a continuous trajectory spanning naive to effector states in CD8 T cells (Fig. 4a). We extracted 3,157 CD8 T cells where we have calculated posterior state probabilities using scChromHMM and have obtained pseudotime estimates ('Reference mapping'). We used the pseudotime values to construct a Gaussian similarity kernel with unit bandwidth, which defines the spatial weights associated with the Moran's I calculation. We calculated Moran's I for each genomic interval as previously described[38] and acknowledge the Trapnell Lab Monocle 3 tutorials for suggesting the use of Moran's I to estimate spatial autocorrelation in single-cell data. We considered all intervals with Moran's $I > 0.15$ as varying across the trajectory. In Fig. 4a, we visualize the posterior probabilities of these regions as a function of pseudotime after applying a moving average filter ($k = 21$).

Using chromVAR to identify associations between motifs and functional states. The chromVAR package[40] identifies associations between transcription factors and chromatin accessibility by examining deviations (i.e., gains/losses) in the accessibility of peaks that share the same DNA sequence motif while carefully controlling for technical biases. Here, we aim to detect associations between transcription factors and the acquisition of distinct chromatin states. We therefore ran the chromVAR algorithm[40], but instead of passing single-cell chromatin accessibility profiles as input, we used our single-cell posterior state probabilities. We repeat this process separately for promoter, enhancer and repressor states. Using the promoter state as an example, chromVAR calculates, for each motif and each cell, the difference between the observed promoter posterior probabilities for peaks containing the motif and the expected posterior probability (based on an average across all cells). chromVAR next performs a similar calculation for a matched background peak set to calculate bias-corrected deviation scores (Fig. 4b and Supplementary Fig. 6c).

Characterizing regulatory priming based on repressive chromatin state. In Fig. 4, we analyze cells based on their posterior probability (repressive state) at 19,306 TSSs. Based on a previous TSS definition[54], we considered all genomic intervals that overlapped with a window starting 2,500 bp upstream of the TSS and extending 500 bp downstream. In each cell, we quantified the level of repression at each TSS as the average of the repressive state posterior probability for each of these windows. We generated a repressive score matrix that represents the level of repression at each cell (rows) at each TSS (columns). We applied TF-IDF normalization[57] to this matrix followed by SVD and used singular vectors 2:30 as input to UMAP visualization in Fig. 4c.

In Fig. 4d, we visualize cell–cell correlations based on the repressive score matrix. In the left heatmap, we use all 19,306 TSS as input to LSI, and use dimensions 2:30 to calculate a correlation matrix. In the right heatmap, we perform the same procedure, but exclude the top 3,000 TSS that were identified as highly variable based on gene expression levels, as determined by SCTransform[59]. In both heatmaps, cells are grouped based on their predicted annotation, but the order of cells within an individual group is random.

In Fig. 4e, we consider the overlap between heterogeneity in repressive chromatin state and transcriptional variation. Using the FindMarkers command in Seurat (default parameters), which performs a Wilcox sum-rank test, we compared CD14 monocytes with all other cells. We performed tests for scChromHMM-derived posterior probabilities (repressive state), H3K27me3 levels

(as measured by scCUT&Tag-pro) and gene expression (as measured by CITE-seq) and kept all results with adjusted $P < 0.01$. To ensure that we robustly identified regions varying in their repressive state, we intersected the lists of TSS identified as differential by scChromHMM and H3K27me3, resulting in 1,597 loci. Of the associated genes, 257 were identified as differentially expressed in the transcriptome.

Comparative analysis to detect differentially expressed genes in the bulk RNA-seq dataset. In Fig. 4e, we show gene loci where we observed cell-type-specific changes in repressive chromatin states and calculated their overlap with transcriptional changes as well. We identified 1,340 loci where we observed shifts in repressive chromatin when comparing CD14 monocytes to remaining cell types, but we did not detect transcriptional differences between these cell types in scRNA-seq data. We further validate the lack in transcriptional shifts of the 1,340 loci using bulk RNA-seq data. We quantify bulk RNA-seq-sorted PBMC dataset using salmon 1.4.0 (ref. [56]) for four major cell types: CD14 monocytes (three replicates), B cells (two replicates), CD8 T cells (two replicates) and natural killer cells (three replicates). We performed differential expression analysis on the bulk RNA-seq data comparing the TPM estimates of the CD14 monocytes with other cell types using DESeq2 (ref. [60]) 1.30.1. Of the 1,340 loci, 1,081 (81%) did not show evidence of differential expression in the bulk RNA-seq dataset (DESeq2 (ref. [60]) adjusted $P > 0.01$). Of the remaining 257 loci that showed transcriptional differences in scRNA-seq data, 74% showed evidence (adjusted $P < 0.01$) in bulk data as well.

Mapping scCUT&Tag datasets to a multimodal reference. In Seurat v4 (ref. [30]), we used sPCA to integrate multimodal (CITE-seq) and unimodal (scRNA-seq) datasets. The sPCA reduction aims to project transcriptomic measurements into a low-dimensional space that most closely resembles the WNN graph. We constructed a WNN graph that identifies, for each cell, a set of neighbor cells in the most similar molecular state based on a weighted combination of RNA and protein modalities. This graph represents a linear kernel, L, that can be used to supervise the dimensional reduction process for transcriptomic measurements.

As described in Barshan et al.[50], there is a closed form solution to sPCA based on a data matrix X, kernel L and centering matrix H. The projection matrix U is represented by the eigenvectors of matrix $XHLHX^T$. We apply the resulting projection matrix (U_{rna}) to new scRNA-seq datasets to identify anchors between query and reference datasets. We can repeat the same process using the protein modality to calculate (U_{ADT}).

Here, we apply a conceptually similar procedure to the integration of multimodal (H3K27me3 scCUT&Tag-pro, from this study) and unimodal datasets (H3K27me3 scCUT&Tag, from Wu et al.[19]). For unsupervised analysis of the scCUT&Tag dataset, we downloaded fragment files from Gene Expression Omnibus with accession code GSE157910. We filtered cells with less than 250 fragments or more than 5,000 fragments, and we repeated the LSI-based analysis workflow to perform unsupervised analysis (Fig. 5a).

For each scCUT&Tag-pro dataset, the reference mapping procedure described above used the protein measurements to map each cell into a reference-defined space. We constructed a k-nearest neighbor graph ($k = 30$) in this space. This defined a kernel matrix L, where $L_{i,j} = 1$ if cells i, j are neighbors and 0 otherwise. Our goal was to compute a SVD of the matrix XL, where X is the TF-IDF normalized peak barcode matrix representing the quantified scCUT&Tag profiles in the multimodal experiment. Unlike scRNA data, the large number of features (peaks) in the scCUT&Tag profiles made it computationally expensive to perform SVD. We circumvented this problem by performing eigendecomposition of the matrix $L^T(X^TX)L$, where the dimensionality is determined by the number of cells but returns the same result. The result was a modified LSI that represents a low-dimensional transformation of the chromatin profiles but has been supervised by the protein measurements. We used the FindTransferAnchors (reduction = 'lsiproject', reference.reduction = 'SLSI', dims = 1:30, k.anchor = 20) to identify anchors between our scCUT&Tag-pro dataset and the unimodal query. We used these anchors to annotate cell types in the query dataset and assign prediction scores.

Lastly, we also used these anchors to interpolate protein values for 173 surface proteins in the scCUT&Tag query dataset. We then applied the previously computed U_{ADT}transformation to map these cells onto the original CITE-seq reference[30], and we filtered cells with prediction score <0.5. We performed this step for visualization purposes only, as it allowed us to view the query scCUT&Tag dataset in the same UMAP visualization as Fig. 2b, which represents all datasets in this study (Fig. 2b).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data generated for this manuscript have been deposited in the Gene Expression Omnibus with accession code GSE195725. The processed datasets are available as open-access downloads at https://zenodo.org/record/5504061. The following public datasets were used in this study: GSM4732109 and GSE164378 (Figs. 2,4 and Supplementary Fig. 3); GSM4732123 (Fig. 1d); GSM1220567, GSM1220569, GSM1027296 and GSM1102793 (Fig. 1g); and GSM5034342 and GSM5034344 (Supplementary Figs. 1c and Fig. 5). Datasets used in differential expression analysis with the bulk total RNA-seq were http://r2platform.com/rna_atlas and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138734.

Code availability

Seurat, Signac and scChromHMM are freely available as open-source software packages at https://github.com/satijalab/seurat, https://github.com/timoast/signac and https://github.com/satijalab/scchromhmm, respectively.

References

56. Srivastava, A., Malik, L., Smith, T., Sudbery, I. & Patro, R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol.* **20**, 65 (2019).
58. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
59. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
60. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

Acknowledgements

We thank all the members of the Satija Lab for thoughtful discussions related to this work. B.Z. is a postdoctoral fellow of the Jane Coffin Childs Memorial Fund for Medical Research. This investigation has been aided by a grant from the Jane Coffin Childs Memorial Fund for Medical Research. This work was supported by the Chan Zuckerberg Initiative (grants EOSS-0000000082 and HCA-A-1704-01895 to R.S.) and the National Institutes of Health (grant K99CA267677-01 to A.S.; grant K99HG011489-01 to T.S.; and grants RM1HG011014-02, 1OT2OD026673-01 and DP2HG009623-01 to R.S.).

Author contributions

B.Z., A.S. and R.S. conceived the study. A.S., B.Z., T.S. and Y.H. performed computational work supervised by R.S.; B.Z., E.M. and I.R. performed experimental work supervised by R.S. All authors participated in interpretation and writing the manuscript.

Competing interests

In the past three years, R.S. has worked as a consultant for Bristol Myers Squibb, Regeneron and Kallyope and served as scientific advisory board member for ImmunAI, Resolve Biosciences, NanoString and the NYC Pandemic Response Lab. P.S. is co-inventor of a patent related to this work. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01250-0.

Correspondence and requests for materials should be addressed to Rahul Satija.

Peer review information *Nature Biotechnology* thanks Barbara Treutlein, Goncalo Castelo-Branco, Anoop Patel and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Rahul Satija

Last updated by author(s): Jan 13, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

The libraries were sequenced on Nextseq550 and demultiplexed using bcl2fastq with the following parameters:

```
bcl2fastq --use-bases-mask=Y34,I8,Y16,Y34 \
  --create-fastq-for-index-reads \
  --minimum-trimmed-read-length=8 \
  --mask-short-adapter-reads=8 \
  --barcode-mismatches 1 \
  --ignore-missing-positions \
  --ignore-missing-controls \
  --ignore-missing-filter \
  --ignore-missing-bcls \
  -r 6 -w 6 \
  -R $fastq \
  --output-dir=$output \
  --interop-dir=$output \
  --sample-sheet=$sheet
```

Data analysis

Custom code: https://github.com/satijalab/scchromhmm

Softwares used in the study:
R packages:
Seurat v4
Signac v1.3.0
chromVAR v1.14.0

ChromHMM v1.23
Alevin v1.4.0
Azimuth tool: azimuth.hubmapconsortium.org
Cellranger-atac v1.2
IGV v2.6.3

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data generated for this manuscript are deposited in the Gene Expression Omnibus with accession code: GSE195725.
The processed datasets are available as open-access downloads at: https://zenodo.org/record/5504061.

Publicly available datasets used in Figure 2, 3, 4 and Supplementary Figure 3:
PBMC ASAP-seq: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4732109
PBMC CITE-seq: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164378

Publicly available datasets used in Figure 1D:
PBMC ASAP-seq: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4732123

Publicly available datasets used in Figure 1G:
CD4 T cells H3K4me1: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1220567
CD8 T cells H3K4me1: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1220569
B cells H3K4me1: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1027296
Monocytes H3K4me1: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1102793

Publicly available datasets used in Supplementary Figure 1C and Figure 5:
PBMC K27me3: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5034342
PBMC K27ac: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5034344

Publicly available datasets used in DE analysis with the bulk total RNA-seq:
http://r2platform.com/rna_atlas
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138734

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We profiled 9,575, 10,386, 8,304, and 8,232 cells for H3K4me2, H3K4me3, H3K9me3 and H3K27me3 histone modification marks respectively on a human PBMC sample; while also generating 12,770 and 15,609 cells each with two replicates for the H3K4me1 and H3K27ac histone modification marks.
Data exclusions	Cells of low data quality were excluded from the analysis (Methods). We filter the fragment file to keep the cellular barcodes (CBs) that were present in the filtered ADT and HTO libraries (10 protein and 10 hashing antibody counts), and also contained at least 100 fragments. Cells that pass these thresholds are used as input to the HTODemux() function in Seurat for demultiplexing and doublet removal.
Replication	Two replicates for each of the H3K4me1 and H3K27ac histone modification marks were generated and the global profiles of each histone mark are reproducible (Supplementary Figure)
Randomization	Randomization is not relevant for this study because there was a single wild type experimental group.
Blinding	Blinding was not relevant for this study because the objective was not to compare two conditions and the data was generated from a single wild type condition.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

The antibodies used were H3K4me1 (1:100, Abcam, ab8895), H3K4me2 (1:100, Abcam, ab32356), H3K4me3 (1:100, Abcam, ab213224), H3K27ac (1:100, Abcam, ab177178), H3K27me3 (1:100, Cell Signaling Technology, 9733), H3K9me3 (1:100, Abcam, ab8898), Phospho-Rpb1 CTD (Ser2/Ser5) (1:50, Cell Signaling, 13546) and guinea pig anti-rabbit (1:100, Novus Biologicals, NBP1-72763). TotalSeq-A conjugated antibodies and panels were obtained from BioLegend (399907, see Supplementary Table 2 for a list of antibodies, clones and barcodes). Fragment Goat Anti-Mouse IgG (Jackson ImmunoResearch, 115-007-003)

Validation

All antibodies are commercially available and validated by the vendor.